A STUDY OF THE RELIABILITY OF TEST QUESTIONS.*

By George Gailey Chambers.

This paper is a study of a test on deductive reasoning and a comparison of the results of that test with the teachers' marks in plane geometry. This test was given to 49 high-school girls who had just completed a half year's work in plane geometry covering during that time the first two books of Robbins & Somerville's text-book. Previous to their study of geometry they had studied algebra through simultaneous quadratics, spending on that 5 periods a week for one school year and 2 periods a week of ¼ of a school year. That was followed by 2 periods a week for ¼ of a school year in constructive geometry. In the case of 5 of these girls we have been unable to obtain the teacher's marks, so that in this paper the answers of only 44 girls are taken into account.

The Association of Teachers of Mathematics of the Middle States and Maryland has a committee at work investigating the results of geometry teaching. A preliminary report of this committee was presented to the Association about a year ago and was published in the recent December issue of the Mathematics Teacher. In this preliminary report it was stated that one of the first tasks that the committee had undertaken was that of preparing suitable non-geometrical tests to determine the results of geometry teaching as to reasoning ability. The writer of this paper happens to be chairman of that committee and the test discussed in this paper is the first one of a series of tests that is being given by the committee as suggested in that report. The purpose in this particular preliminary work is to determine the reliability of the questions used, rather than to test the results of geometry teaching. I am giving this discussion to this body now to inform them as to the progress of the work of the com-

* Read before the joint meeting of the New England Association of Mathematics Teachers and the Association of Teachers of Mathematics of the Middle States and Maryland, February 28, 1914.

mittee and especially to enable the committee to obtain frank criticisms of its work as it progresses. I should say, however, that it is not a report of the committee, and that any opinions expressed in this paper are opinions of the writer only.

One of the most pertinent problems before the educational world today is that of measuring results of educational processes. This general problem has been studied by a number of investigators and it has led to the introduction into educational literature of several technical terms such as, tables of distribution, frequency curves and coefficients of correlation. There is much literature on the subject but very little of it in those periodicals that come into the hands of mathematics teachers in general. It seemed wise therefore to present here a concrete example illustrating the meaning of some of the technical terms now current.

This is the set of questions used:

I. Do you discover any defects in the following reasoning, and if so, explain why it is defective?

The sidewalk was wet this morning. Therefore it must have rained last night.

II. If all the inhabitants of the Rahib Islands have blue tattoo marks on their bodies, then which of the following statements would necessarily be true, which could not be true, and which might possibly be true?

(1) All people who have blue tattoo marks on their bodies are inhabitants of the Rahib Islands.

(2) Some inhabitants of the Rahib Islands do not have blue tattoo marks on their bodies.

(3) No people with blue tattoo marks on their bodies live anywhere except on the Rahib Islands.

(4) Some of the inhabitants of the Rahib Islands have blue tattoo marks on their bodies.

III. A certain club wishes to select the evening for its regular weekly meeting which would be most satisfactory to its members. Accordingly the secretary wrote to each member, asking what evening would be most satisfactory.

Can you suggest another question which would have been better for the secretary to have asked?

IV. If a photographic plate be exposed to X-rays and then developed, black marks will be found upon it.

If upon developing a photographic plate you should find black marks upon it what would you conclude?

Also if you should not find black marks upon it what would you conclude?

V. If John agrees to join the football team provided Charles joins it, but Charles decides not to join it, what follows about John? If John joins, but Charles does not join, is John breaking his agreement?

A multigraphed copy of these questions was put into the hands of each pupil. No instructions were given except the following which was placed at the head of the multigraphed copy:

"Please write answers to the following questions, first reading them all before you begin writing. Answer them in any order you wish, numbering your answers to correspond to the numbers of the questions."

This test was given under the charge of two English teachers in the school and particular care was taken to see that there was nothing said or done to indicate in the minds of the pupils that there was any connection between this test and their work in mathematics. No inquiries in regard to the meaning of the questions were answered.

The answers to these questions were read carefully to discover the specific acts of correct deductive reasoning which appeared in the answers. The largest number of points made by any one pupil was 8, and the smallest number was 1, a point being scored for each act of correct deductive reasoning.

I will now give a resumé of the most common answers to the various questions, and also some significant but less common answers:

Question I.: Thirty-one stated that the sidewalk might have been wet from other causes and that therefore the reasoning was defective. A point was scored for each of these.

Three dealt with criticisms of the use of the pronoun "it" in the statement of the problem, declaring that it referred either to the sidewalk or to the morning, and that it was in that respect that the statements were defective.

Three stated that the two statements were defective because they gave the effect before the cause.

Two specifically declared that the reasoning was correct as given.

Question II. (1): Sixteen stated that it might possibly be true, or it is not necessarily true. A point was scored for each of these.

Six that it is not true because there are many people not inhabitants of the Rahib Islands who have blue tattoo marks on their bodies. From the evident premise in the mind of the pupil the reasoning is correct and a point was scored for each. These answers, however, show that this question is defective in that it is so easy for the pupil to add another premise.

Four that it is not true because some people not living on the Rahib Islands might have blue tattoo marks. This answer in itself is illogical.

Ten that it was necessarily true.

Eight that it could not be true.

Question II. (2): Thirty-one answered that this statement could not be true. A point was scored for each of these.

Three that it might possibly be true because foreigners visiting these islands would not have blue tattoo marks. Evidently the given statement was interpreted loosely in these cases; that is, that there were probably exceptions to it. With that interpretation the reasoning was logical and these questions were counted as correct. The defect was in the looseness of the interpretation and not in the reasoning.

Three that it might possibly be true because there may be a few who do not have blue tattoo marks, by being accidentally missed or by not believing in the "severe custom." In these cases also the pupils recognized that the true facts were contradictory if the two statements were taken as holding without exception. Evidently the reasoning from the premises in the minds of the pupils was correct.

Five that it might possibly be true without any indication as to any modification of the given premise. These were not counted as correct.

One that it would necessarily be true.

Question II. (3): Twenty-four replied that this statement might possibly be true. A point was scored for each of these.

Two that it is true.

Eight that it could not be true, but giving an explanation showing that they had added an additional premise to the given

one. Their reasoning was evidently correct from the premises in mind, and a point was scored for each.

Six that it could not be true, but giving no indication of any added premise.

Question II. (4): Nineteen replied that it is necessarily true. A point was scored for each of these.

Two that it might possibly be true.

Nine that it is not true because it implies that some of the inhabitants do not have blue tattoo marks. With this interpretation of the statement the logic is correct and a point was scored for each.

Ten that it is not true but giving no indication of their interpretation of the given statement.

Question III.: Five suggested that he might better have asked what evenings were not satisfactory. A point was scored for each.

One that he might have asked what evenings were open. A point was scored for this.

Twenty-one that he should have named one or possibly two evenings and asked the members whether they suited. These were not counted as correct.

One that the secretary should have asked the members to come out at an appointed time and then decide the evening for the meeting. This was not counted.

Eight suggested other words instead of the words "most satisfactory," but made no suggestions affecting the reasoning involved. These were not counted.

Question IV. (1): Twenty-eight concluded that the plate had been exposed to X-rays. These were not counted as correct.

Five recognized that there might have been other causes for the black marks. A point was scored for each of these.

Question IV. (2): Thirty-two concluded that the plate had not been exposed to X-rays. A point was scored for each of these.

Five that it had not been exposed to X-rays nor to anything else that might have caused black marks. A point was scored for each of these.

Question V.: Six answered that John would not be breaking his agreement. A point was scored for each.

Twenty-eight that John would be breaking his agreement. These were not counted as correct.

One of these stated that it should not be considered as the breaking of an agreement because the matter was not of such importance that to change his mind could be called breaking his agreement.

Another stated: "If John joins the football team but Charles does not join, then John is breaking his agreement provided the agreement was written and signed by both parties; if not, John is not doing anything wrong because he merely agrees to that and is eligible to break his word."

As far as the logic is concerned these last two answers are not different from the other six; but they are interesting as indicating the moral point of view of the pupil.

This leads to the following summary giving the number of pupils who scored on the various questions considering each of the four parts of question II. as independent questions, and likewise the two parts of question IV.

TABLE I.

Question.	No. Scoring.
I	31
II (1)	22
II (2)	37
II (3)	32
II (4)	28
III	6
IV (1)	5
IV (2)	37
V	7

No two pupils scored on both question III. and question IV. (1). Three scored on both question IV. (1) and question V. One scored on both question III. and question V.

These results may also be arranged as in the first two columns of the following table. Those two columns of this table constitute what is called a table of distribution.

They may also be represented by a frequency curve, as in Fig. 1.

This is drawn on the assumption that the difference in the amount of the ability in question between a pupil who made, for example, two points and a pupil who made one point is equal to the difference in the amounts between two pupils scoring any other two consecutive numbers.

TABLE II.

Number of Points Scored by One Pupil.	Number of Pupils Making the Corresponding Score.	Rank.
8	1	1
7	3	3
6	11	10
5	11	21
4	5	29
3	10	36½
2	2	42½
1	1	44

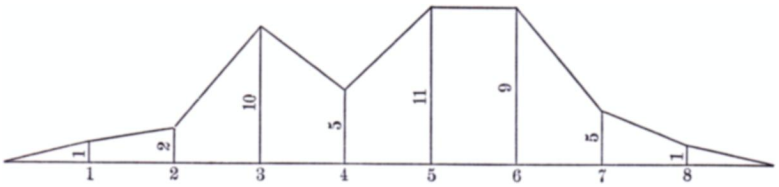

FIG. 1.

Table III. contains a similar table of distribution for the same 44 pupils based upon the teacher's mark in geometry.

TABLE III.

The Teacher's Marks.	Number of Pupils Receiving the Corresponding Mark.	Rank.
A	5	3
B+	5	8
B	11	16
C+	17	30
C	6	41½

A is the highest mark.

Fig. 2 is the corresponding frequency curve.

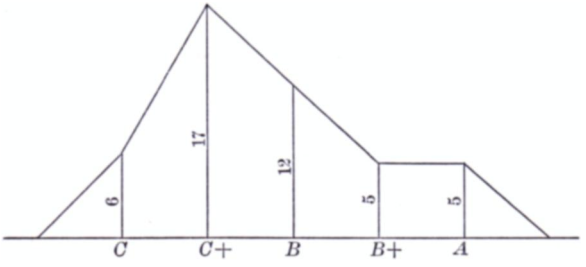

FIG. 2.

Theoretically a frequency curve should take the form shown in Fig. 3.

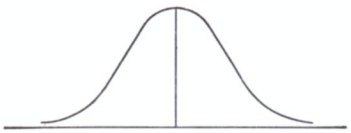

FIG. 3.

On comparing Fig. 1 with Fig. 3 we observe that the number of pupils scoring 4 points is abnormally small, but that otherwise the curve of Fig. 1 approximately coincides with a curve of the type shown in Fig. 3.

A similar comparison in the case of Fig. 2 shows that the number of pupils marked C+ is abnormally large.

We will now study the correlation between the series of measures obtained from these test questions and the series of measures given as teacher's marks; that is, roughly speaking, the tendency that a pupil having a high measure in one series will also have a high measure in the other series; a low measure in one, a low measure in the other; and a medium measure in one, a medium measure in the other.

Since the two sets of measures are so essentially different, we must make use of ranks, that is, the numbers which give the positions of the pupils when they are arranged in a series according to merit. Referring to Table II., it is evident that the pupil making eight points should have the rank 1. It is also evident that the three pupils making 7 points each should have the ranks from 2 to 4, but we have no way of assigning these ranks to individual pupils. The best we can do is to assign to each the same rank. In this and the following cases the average of the real ranks has been assigned in Table II. In the same way the ranks have been assigned in Table III.

This leads to the following table giving the 44 pupils, with their corresponding ranks in the two series.

The mathematical theory of probability gives the formula

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)},$$

where d is the difference between the ranks of the same pupil in the two series, Σd^2 is the sum of the squares of those differences,

n is the number of pupils, and ρ is a measure of the correlation. To avoid fractions I have written the formula as follows:

$$\rho = 1 - \frac{3\Sigma(2d)^2}{2n(n^2 - 1)}.$$

The data of Table IV. gives $\rho = .48$.

TABLE IV.

Pupil.	Ranks from Teacher's Marks in Geometry.	Ranks from Results of This Test.	Differences in Ranks.	Squares of Twice the Differences.
E. B.	30	10	20	1,600
M. B.	30	36½	6½	169
R. B.	16	21	5	100
L. S. B.	8	3	5	100
M. E. B.	41½	21	20½	1,681
I. M. C.	30	29	1	4
M. J. C.	41½	29	12½	625
M. C.	30	36½	6½	169
H. A. D.	3	10	7	196
L. D.	16	10	6	144
M. D.	30	29	1	4
M. E.	30	36½	6½	169
M. F.	16	21	5	100
J. B. F.	30	21	9	324
J. F.	8	10	2	16
I. T. G.	16	10	6	144
F. H.	41½	29	12½	625
T. H.	41½	29	12½	625
S. J.	41½	36½	5	100
E. V. J.	16	10	6	144
M. M. L.	8	21	13	676
H. W. L.	16	21	5	100
R. G. L.	41½	44	2½	25
M. M.	16	3	13	676
E. M.	3	36½	33½	4,489
E. T. M.	8	10	2	16
E. M. M.	30	36½	6½	169
D. M.	30	21	9	324
R. N.	30	36½	6½	169
E. N.	3	10	7	196
E. O.	8	1	7	196
C. P.	30	21	9	324
A. P.	16	42½	26½	2,809
E. P.	16	36½	20½	1,681
O. R.	3	10	7	196
E. R.	30	10	20	1,600
G. R.	30	10	20	1,600
M. R.	3	21	18	1,296
H. M. S.	16	21	5	100
F. S.	16	36½	20½	1,681
D. V. S.	30	3	27	2,916
E. Y.	30	21	9	324
R. Z.	30	42½	12½	625
P. Z.	30	36½	6½	169

The use of ranks instead of actual measures has introduced an error. This can be at least partly corrected by the formula

$$r = 2 \sin\left(\frac{\pi}{6}\rho\right).$$

This gives $r = .50$, as the corrected coefficient of correlation.

A natural question to ask is as to the probability that the apparent correlation in this case is a real correlation due to a functional relation rather than to pure chance. This is answered by determining the probable error. Recourse again to the theory of probability shows that the probable error of r is given by the expression:

$$0.706\frac{1 - r^2}{\sqrt{n}}.$$

In this case the probable error is .08. That means that it is an even chance that the true value of r is between $.50 - .08$ and $.50 + .08$, that is, between .42 and .58.

The chances are 16 to 1 against the true value of r differing from the above value, .50, by more than three times the probable error; that is, the chances are 16 to 1 that the true value of r is between .28 and .74.

Similarly the chances are 1,000 to 1 that the true value of r is between $.50 - 5 \times .08$ and $.50 + 5 \times .08$; that is, between .10 and .90.

It is true that the fact that there were so many groups of pupils with the same rank may increase the probable error, yet there is reason to think that notwithstanding that fact, the chances are large in favor of the true value of r being greater than .3. That means that we can feel reasonably certain that the traits measured by these two sets of measures are correlated; that is, as expressed before, there is a tendency for a pupil high in one series to be high in the other, and so forth.

As stated in the beginning, one purpose before me has been to give this body of teachers a concrete illustration of some of the applications of the theory of probability to experimental educational data. May I suggest that a very valuable course in any teacher's preparation is a course in the theory of probability with special application to educational statistics. Such a course is especially valuable to a teacher of mathematics.

My main purpose, however, was to determine whether this set of questions based on non-mathematical subject matter had any real value in measuring reasoning ability. The conclusion seems to be fully justified that it has a value for that purpose. It is also evident that some of the questions can be modified so that the set will serve as a still better means for that purpose. This has an important bearing on the general question of testing the results of geometry teaching, because such a test can be given to groups of individuals some of whom have not studied geometry, while others have studied that subject. The results can then be investigated to determine whether or not those who have studied geometry tend to rank higher than those who have not studied that subject.

Note.—Before reading this paper I gave opportunity for the persons present at the meeting to answer the same set of questions. Forty-eight persons answered them. I have marked these answers in the same way as I marked the answers of the pupils, with the following results:

Question I.: Forty-five stated that the sidewalk might have been wet from other causes and that therefore the reasoning was defective.

Question II. (1): Forty-five stated that it might possibly be true or it is not necessarily true, one that it was necessarily true, two that it could not be true.

Question II. (2): Forty-eight answered that this statement could not be true.

Question II. (3): Forty-six answered that this statement might possibly be true, two that it could not be true.

Question II. (4): Forty-four replied that it is necessarily true, one that it might possibly be true, two that it is not true.

Question III.: Twenty-five suggested that he might better have asked what evenings were not satisfactory. Ten that he might have asked what evenings were satisfactory. One that he should have named one or possibly two evenings and asked the members whether they suited.

Question IV. (1): Six concluded that the plate had been exposed to X-rays. Thirty-six recognized that there might have been other causes for the black marks.

Question IV. (2): Thirty-four concluded that the plate had not been exposed to X-rays.

Question V.: Twenty-eight answered that John would not be breaking his agreement. Seven that John would be breaking his agreement.

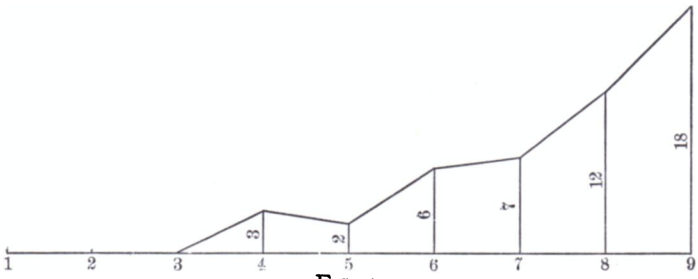

FIG. 4.

The following table gives a summary of the number of persons who scored on the various questions. This corresponds to Table I. of the above paper.

TABLE V.

Question.	No. Scoring.
I	45
II (1)	47
II (2)	48
II (3)	46
II (4)	46
III	35
IV (1)	36
IV (2)	34
V	28

These results also lead to the following distribution table which corresponds to Table II. above.

TABLE VI.

No. of Points Scored by One Person.	No. of Persons Making the Corresponding Score.
9	18
8	12
7	7
6	6
5	2
4	3

The corresponding frequency curve is shown in Fig. 4, which corresponds to Fig. 1 above.